DOCUMENT RESUME

ED 137 486 UD 016 911

AUTHOR Lord, Frederic M.
TITLE A Study of Item Bias Using Characteristic Curve Theory.
SPONS AGENCY College Entrance Examination Board, New York, N.Y.
PUB DATE Jul 76
NOTE 17p.; Graphs may be marginally legible due to small print

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Tests; *Black Students; *Caucasian Students; Comparative Analysis; Educational Testing; *Item Analysis; *Test Bias; *Test Interpretation; *Verbal Tests
IDENTIFIERS *Scholastic Aptitude Test

ABSTRACT

This study investigates whether item characteristic curves are the same for black students as for white students in the United States. The data analyzed were the answer sheets of 2269 black students and 2285 white students taking the 85-item Verbal Section of the College Board's Scholastic Aptitude Test. The study of item characteristic curves is a feasible and fruitful way to investigate item biases. It has definite advantages over less sophisticated methods. More than a third of the 85 test items were found to have different characteristic curves for blacks and for whites at the 5% level of statistical significance. It is in many cases not clear from reading it why a particular item is biased in a particular way. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED137486

Final to Proceedings 7/31/76

A Study of Item Bias, Using Item Characteristic Curve Theory

Frederic M. Lord

Educational Testing Service

Princeton, New Jersey 08540, U.S.A.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

July 1976

VD016911

A Study of Item Bias, Using Item Characteristic Curve Theory

Frederic M. Lord
Educational Testing Service
Princeton, New Jersey 08540, U.S.A.

I am going to report on a study of bias in test items. The study compares data from about 2250 whites with data from an equal number of blacks. Both groups are about 44 percent male. The test administered is an 85-item verbal test used for college admissions--the Verbal Section of the April 1975 Scholastic Aptitude Test of the College Entrance Examination Board. There are four kinds of verbal items in the test: verbal analogies, antonyms, word meaning, and reading comprehension.

Does this test measure the same thing for blacks as it does for whites? Are there some items that should be removed from the test so that the remaining items will measure appropriately in both groups? These are the questions that we are trying to answer.

The general plan and design of the study was developed by Gary Marco, Director, Statistical Analysis, College Board Programs Division, at Educational Testing Service. Marco will be the senior author of the final report of this study. The study is partially supported by the CEEB. Before giving more details, I will talk about certain previous approaches to the study of item bias, also about item characteristic curve theory, upon which the present study is based.

Figure 1 plots item difficulty for blacks against item difficulty for whites. For the present, I use the term 'item difficulty' to refer to the proportion of correct answers given to an item. The data used to obtain Figure 1 are the same data already described. The 85 crosses in the figure represent the 85 items in the verbal test. Items falling along

the dashed line in the figure are items that are as easy for blacks as for whites. Items below this line are easier for whites. The solid oblique line is a straight line fitted to the scatter of points. The solid line differs from the dashed line because whites score higher on the test than blacks. If all the items fell directly on the solid line, we could say that the items are all equally biased; or, conceivably, equally unbiased.

It has been customary to look at the scatter of items about the solid line and to pick out the items lying relatively far from the line and consider them as atypical and undesirable. In the middle of Figure 1 there is one item lying far below the line that appears to be strongly biased in favor of whites; also another item far above the line that favors blacks much more than other items. A common judgment would be that both of these items should be removed from the test.

In Figure 1 the standard error of a single proportion is about .01, or less. Thus most of the scattering of points is not attributable to sampling fluctuations. Unfortunately, the failure to fall along a straight line is not <u>necessarily</u> attributable to differences among items in bias. This is true for six different reasons, which I will discuss next.

In the first place, we should expect the scatter in Figure 1 to fall along a curved line, not a straight line. Logically, the curved line must pass through the points (0,0) and (1,1). If the groups performed equally

well on the test, the points could fall along the dashed line; but since one group performs better than the other, most of the points must lie to one side of the dashed line and the relationship must be curved.

Careful studies attempt to avoid this curvature by transforming the proportions. If an analysis of variance is to be done, the conventional transformation is the arcsine transformation. The real purpose of the arcsine transformation is to equalize sampling variance. Whatever effect it may have in straightening the line of relationship is purely incidental.

The transformation usually used to straighten the line of relationship is the inverse normal transformation. The proportion of correct answers is replaced by the relative deviate that would cut off the same proportion of the area under the standard normal curve. The result of this transformation is shown in Figure 2. Indeed, the points in Figure 2 fall about a line that is more nearly straight than was the case in Figure 1.

Unfortunately, there are theoretical objections to the inverse normal transformation. Suppose that the test were to contain several items so difficult that everyone simply guessed at random on these items. Since the items here are five-choice items, the proportions of correct answers for both blacks and whites would be .20. This means that the curve in Figure 2 should pass through the point (-1.84, -1.84). It again appears that when there is guessing, the points in Figure 2 cannot be expected to fall strictly along a straight line unless the two groups perform equally well on the test.

Next, there is a reason why the items cannot be expected all to fall along a single curve. If items at one level of discriminating power fall along a certain curve, then items at a different level of discriminating power will fall along a different curve. The reason is that the more discriminating items would produce more difference between blacks and whites than would the less discriminating items.

This last leads to the startling conclusion that the proportion of correct answers really is not a measure of item difficulty! Let me come back to this point in a moment.

Figure 3 shows some typical item characteristic curves. The scale along the baseline represents the ability of the examinee. The item characteristic curve shows the probability of a correct answer as a function of examinee ability. The general shape of the curve follows naturally from the fact that success on the item tends to increase with ability, but the probability of success can never exceed 1.0, nor fall below 0. Such curves typically have one point of inflection.

In item characteristic curve theory it is usually assumed that such curves can be defined by three item parameters. The item difficulty b represents the ability level corresponding to the point of inflection. When there is no guessing, b is the ability level at which the examinee has a 50 percent chance of answering the item correctly. The higher the value of b , the more difficult the item.

The slope at the inflection point is proportional to the item parameter a , which represents the discriminating power of the item. When there is no guessing, the slope at the point of inflection under a commonly used model is simply $a/\sqrt{2\pi}$.

The item parameter c represents the probability of success for examinees of infinitely low ability. Thus c defines the lower asymptote of the item characteristic curve. It is nonzero whenever examinees can guess the correct answer. Typically, but not always, c is less than the chance level that would be achieved by an examinee guessing at random. The reason is that test developers spend much effort and ingenuity providing attractive distractors to the items, with the result that people who do not know the answer typically do less well than if they had chosen their responses at random.

Figure 4 shows two rather different item characteristic curves; inverted on the baseline are the distributions of ability for two different groups of examinees. First of all you should note: The item difficulty b should be the same regardless of the group from which it is determined; the ability required for a certain level of performance by an individual does not depend on the ability distribution of other people in some group. The same holds true for the slope a at the inflection point, and for the lower asymptote c . This invariance is the outstanding advantage of the item parameters used in item characteristic curve theory. In principle, within reasonable limits, the parameters should stay the same regardless of the group tested.

Now please note carefully the following. In group A , item 1 is answered correctly less often than item 2. In group B , the opposite occurs. If we use the proportion of correct answers as a measure of item difficulty, we find that item 1 is easier than item 2 for one group, but harder than 2 for the other group. It is for this reason that I assert

that proportion of correct answers in a group of examinees is not a measure of item difficulty.

This proportion not only describes the test item but also describes the group tested. This is a basic objection to analyzing item bias by the approaches suggested by Figures 1 and 2.

Still another difficulty with these conventional approaches may be mentioned. The black group and the white group represented in Figure 1 are apparently not comparable in verbal skills. It might be argued that we should base our analysis on white and black groups that are matched on verbal skills. Such matching is difficult to carry out in practice, however. We cannot properly match on a test composed of the items that are to be studied, since this would introduce spurious relationships. If we try to match on a parallel form of the same test, we will be matching on a fallible score when we should be matching on a true score. There will be a regression effect that will present proper matching.

One way to compare the performance of blacks and whites at the same level of verbal skill is to compare the characteristic curve of an item for blacks with the characteristic curve of the same item for whites. Any difference between the curves indicates some kind of bias. This comparison is made in the study I am reporting today.

Before proceeding let me note the following, however. Suppose, to take an extreme example, certain items in a test are taught to one group

of students and not taught to another, while other items are taught to both groups. This way of teaching increases the dimensionality of whatever is measured by the test. If the items would otherwise have been factorially unidimensional, this way of teaching will introduce additional dimensions. If we ignore this and analyze all items as if they were unidimensional, we cannot expect all item characteristic curves to be the same for both groups. Since blacks and whites are exposed to different learning environments, the situation may be quite similar for them. With this in mind, let us turn to a report of the present study.

We used a computer program, LOGIST, which simultaneously estimates the ability of each examinee and the a, b, and c parameters of each item. The answer sheets of the 2250 whites and the answer sheets of the 2250 blacks were first run separately on this program.

It is inherent in the nature of the problem that the origin and the unit for measuring ability cannot be determined from the data. Thus the item parameters from the black group cannot be compared directly with the item parameters from the white group. To determine a common origin and unit, we plotted the b parameters (item difficulties) for the black group against the b paremeters for the white group. The plot is shown as Figure 5. This plot is the same as Figures 1 and 2 except that here item difficulty is measured by the parameter b.

According to the icc model the values of b for blacks and for whites can only differ in origin and unit of measurement. The straight line fitted to the 85 points is the first principal axis. This line was used to put all item parameters on the same scale.

We can now test the null hypothesis that a particular item has the same item characteristic curve for blacks and for whites.* The asymptotic significance test used will be discussed in a moment. Forty-six of the 85 items were found to be significantly different at the five percent level.

The study could have been stopped at this point. However, it might be argued that a test composed of so many biased items did not provide an adequate basis for measuring examinee ability. To meet this objection, the items showing significant difference beyond the 15 percent level were eliminated, leaving 32 items for which the black and white item characteristic curves were very similar.

The black and white groups were now combined and the data for the 32-item test run on LOGIST, ignoring color differences. In this way, the ability parameters of blacks and whites on the 32-item test were all estimated on the same scale.

As a final step, the entire first step of the study was repeated, now treating the ability parameters just estimated as given. Since the ability parameters are all on the same scale, the item parameters obtained for the black group are now comparable with the item parameters obtained for the white group.

Asymptotic significance tests were again carried out to test the null hypothesis that for a given item the black and the white item characteristic

---

*Actually, in order to make a significance test possible, the value of the c parameter for an item was required to be the same for blacks and for whites. Thus the curves could only differ in a and b parameters. This complication is glossed over here but will be fully covered in the final report.

curves are identical. This hypothesis was rejected at the 5 percent level for 38 items out of 85. The distribution of the 85 items over different significance levels is shown in Table 1.

I should now discuss the rationale for the significance tests. Actually, it is not presently possible to specify with certainty even the asymptotic standard error of the maximum likelihood estimates used in this study. An approximation, based on certain reasonable assumptions to these standard errors was used. Rather than trying to justify the approximation mathematically, it may be more satisfactory to justify it by the results of an empirical study carried out especially for this purpose, as follows.

The black and white groups were combined into a total group of about 4500 individuals. This total group was divided at random into two groups, which we may designate 'blue' and 'red.' The entire statistical analysis involving at least three LOGIST runs was repeated for these two random groups. At the end, asymptotic significance tests were carried out to test the null hypothesis that the blue item characteristic curves were the same as the red. The distribution of the 85 test items over various significance levels is shown in Table 2.

Since the blue and the red groups were drawn at random, the 85 items should be rectangularly distributed over the range of significance levels. This would mean just eight and one-half items in each probability interval of width .10. The frequencies shown in Table 2 are surprisingly close to this, suggesting that the statistical procedure used is actually a good approximation. When we compare Tables 1 and 2, it seems that a third

or more of the items really have different characteristic curves for blacks and for whites.

Figure 6 illustrates one such curve--the curve for item 71. The base line of the figure represents examinee ability over the range from -4 standard deviations to +3 standard deviations. The vertical axis shows the probability of a correct answer to item 71. The dashed curve is the icc for blacks; the solid curve is the icc for whites. At the extremes of each group, individuals are shown as points, in order to give an idea of where the data lie. In the middle of the curve, where most of the data lie, individual points are not shown. It should be remembered that a particular individual in practice answers an item either correctly or incorrectly--we do not actually observe a probability for a single individual.

The two curves for item 71 are significantly different beyond the .01 level. Interestingly, high-level white students do better than high-level black students on this item, but low-level black students do better than low-level white students.

A similar situation appears in the next figure which shows the results for item 2. In addition, we find that item 2 does not discriminate among black students but does discriminate among white students.

Item 71 and item 2 illustrate a kind of difference that would not be found by the techniques shown in Figure 1 and Figure 2. Each of these items falls along the curve of relation in Figures 1 and 2 and does not appear to be more difficult over all for one group or the other.

The last figure shows the item characteristic curves for item 24, which is a very difficult item. Regardless of ability levels, black students are unsuccessful on this item. For white students, however, the item does discriminate at very high ability levels.

I have studied the items in the test and compared them with the statistical results without reaching any startling insight into the reasons for the special biases of individual items. Unfortunately, I cannot hand out a copy of all the test questions together with the table of the statistical results for you to study. The reason is that the items we have analyzed are still in our active item pool for use in building new college admissions tests. The items together with the past statistical analyses are expensive, and the confidentiality of the items must be maintained. I have permission to read to you the three items represented by the last three illustrations. Perhaps you will see some clear explanation for the statistical results.

71. A deficiency of calories means a shortage of the supply of calories to the body in relation to the ---- them.

(A) production of
(B) variations between
(C) assessment of
(D) requirement for
(E) connections among

2. INJURE: (A) release (B) refrain
(C) smooth (D) embellish (E) heal

24. We do not have a full grasp on experience until we have symbolized it; we cannot ____ until we have ____ .

(A) understand .. learned
(B) communicate .. thought
(C) inform .. revised
(D) explain .. hypothesized
(E) know .. verbalized

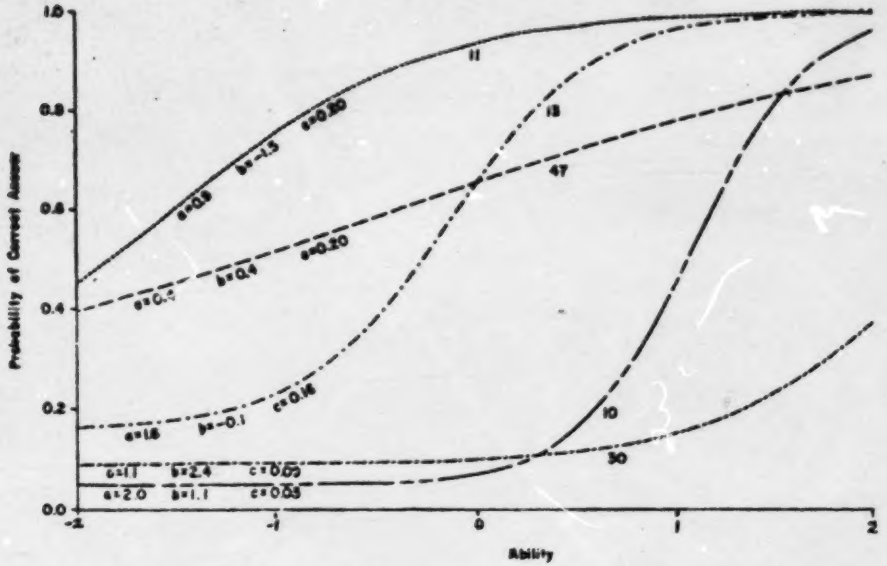
The final report of this study will include not only the material I have presented here, but also, for comparison, the statistical analysis of the same data by the method illustrated in Figure 2. A more thorough study of the items at that time may reveal more clearly the reasons for the biases shown.

Does the test measure the same psychological trait for blacks as for whites? If it measured totally different traits for blacks and for whites, the scatterplot in Figure 5 would show little or no relationship between the item difficulty indices for the two groups.

In view of this, the study shows that the test does measure approximately the same skill for blacks and whites. Some items show up differently in the two groups, but the differences are rather small.

The item characteristic curve techniques used here can pick out certain atypical items that should be cut out from the test. It is to be hoped that more careful study of such analyses will help us understand better why certain items are biased, why certain groups of people respond differently than others on certain items, and what can be done about this.

Fig. 3. Probability of correct answer as a function of ability, as estimated for five SAT Verbal items.



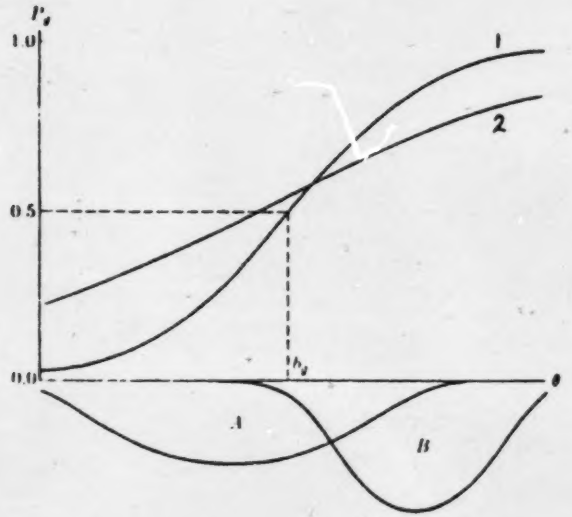


Fig. 4. Item characteristic curves in relation to two groups of examinees.

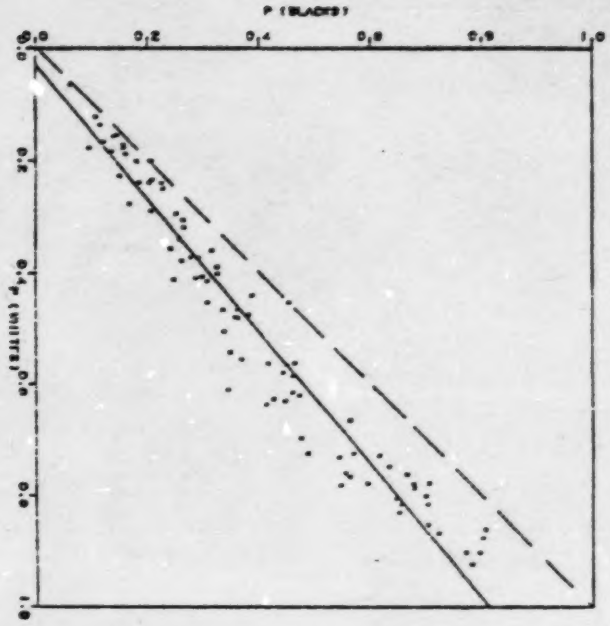


Fig. 1. Proportion of right answers to 85 items, for blacks and for whites.

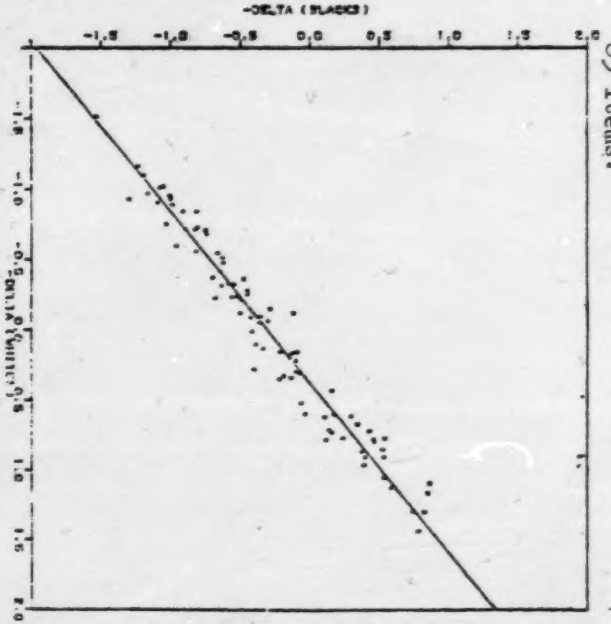


Fig. 2 Inverse normal transformation of proportion of correct answers for 85 items.

Fig. 5. Difficulty parameters (b) for 85 items for blacks and for whites.

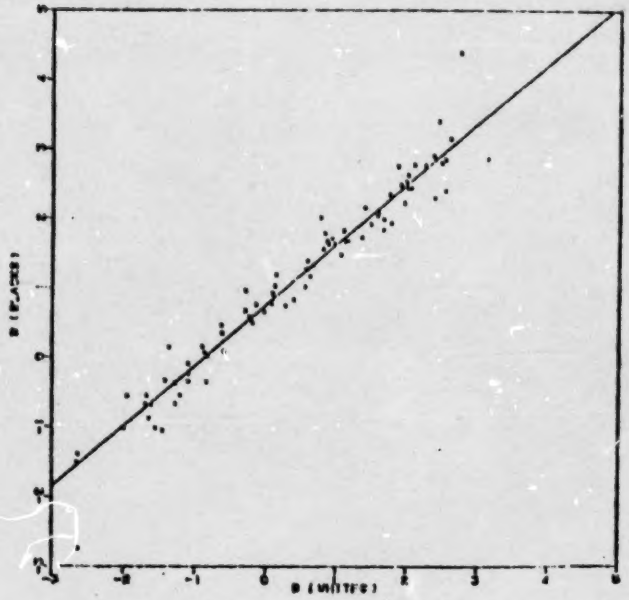


Fig. 6. Black (dashed) and white (solid) characteristic curves for item 71.

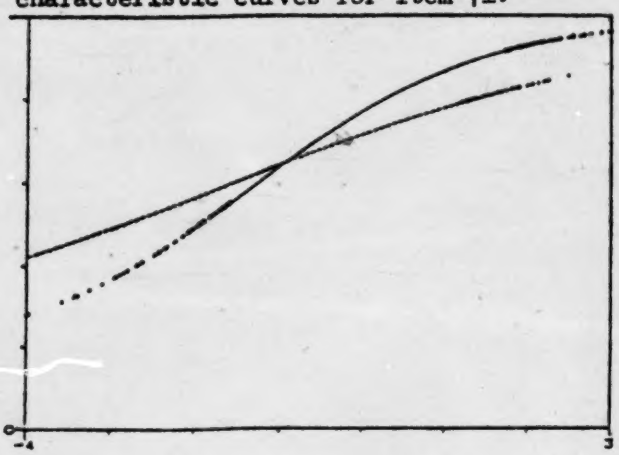


Fig. 7. Characteristic curves for item 2.

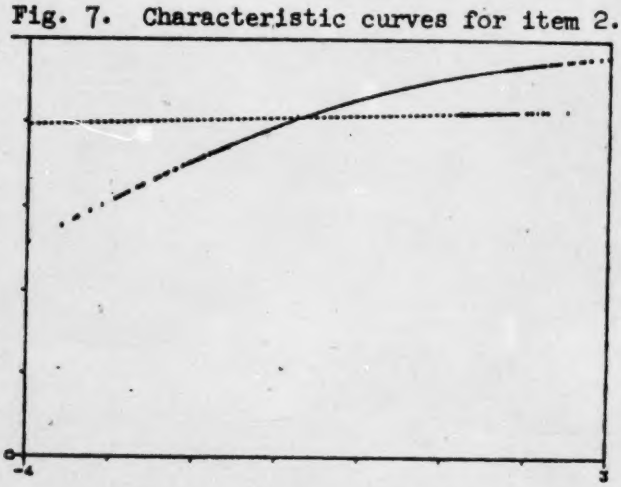


Table 1

Significance Level	No. of Items
.00 - .05	38
.05 - .10	4 } 42
.10 - .20	3
.20 - .30	3
.30 - .40	3
.40 - .50	8
.50 - .60	4
.60 - .70	9
.70 - .80	3
.80 - .90	6
.90 - 1.00	4

Table 2

Significance Level	No. of Items
.00 - .05	3
.05 - .10	6 } 9
.10 - .20	6
.20 - .30	6
.30 - .40	11
.40 - .50	9
.50 - .60	7
.60 - .70	5
.70 - .80	12
.80 - .90	10
.90 - 1.00	10

Fig. 8. Characteristic curves for item 24.